Winning Space Race with Data Science

Bryan Moncada
04/23

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

EXECUTIVE SUMMARY

SpaceX cost for Falcon 9 rocket launches is only 62 million dollars, while the competition cost is 165 million. This saving due to the reason that SpaceX reuses the first stage. In this project, we want to determine the cost of a launch by finding if a launch is successful or not. The way we obtain launch success is by obtaining data from the SpaceX API and Web Scraping to later train a classification model to predict if the next launch will land or not. Our methodologies include EDA with matplotlib and SQL, and interactive data visualization with Folium and Plotly. The resulting model is a KNN with 0.94 accuracy that will let us predict with acceptable confidence the success or failure of the next launch.

INTRODUCTION

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars. Other providers cost upward of 165 million dollars each.
- Much of the savings is because Space X can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch
- This information can be used if an alternate company wants to bid against space X for a rocket launch.
- We create a machine learning pipeline to predict if the first stage will land given some data.
- The data is collected from the SpaceX API and Wikipedia Web Scraping

Section I

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX API + Web Scraping

Perform data wrangling

- Identify missing values, convert landing outcomes to training labels

Perform exploratory data analysis (EDA) using visualization and SQL

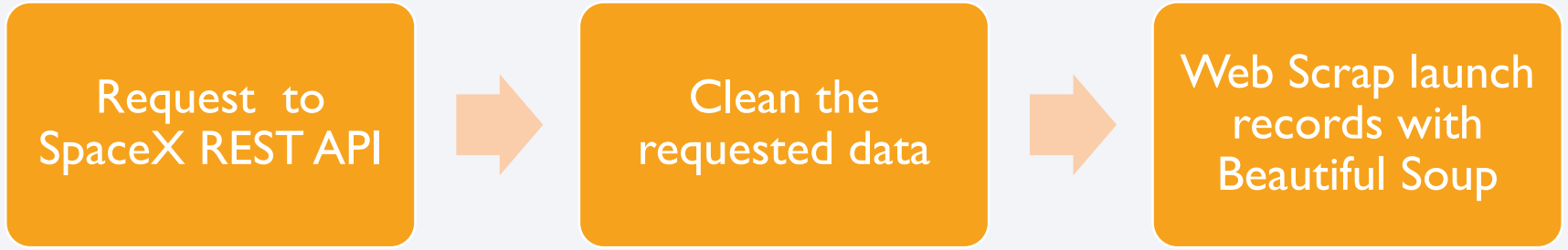
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Train data in collected data, find the best model that has highest accuracy

Data Collection

- SpaceX data collected from SpaceX REST API
- Web scraping Falcon 9 from Wikipedia



Data Collection – SpaceX API

- Request to SpaceX API and Clean Data

- Notebook Link to GitHub

Request and Parse the SpaceX API

Filter Data Frame by Falcon 9

Data Wrangling

- Missing Values

Data Collection - Scraping

- Perform web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled: “List of Falcon 9 and Falcon Heavy launches"
- Notebook Link to GitHub

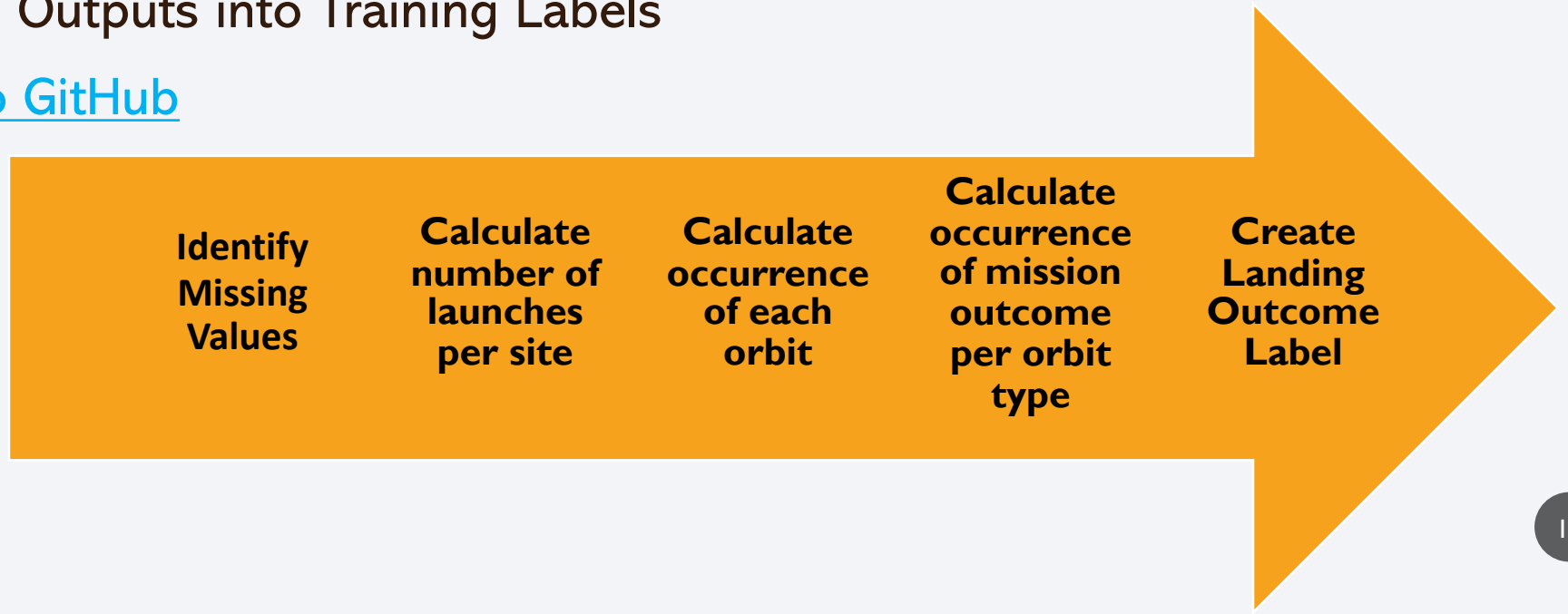
Request Falcon9 Launch Wiki from URL

Collect relevant column names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Find patterns in the data and determine the outcome label
- Different Landing Outputs:
  - `True Ocean / False Ocean`
  - `True RTLS / False RTLS.  (RLTS -> landed to a ground pad)`
  - `True ASDS / False ASDS.  (ASDS -> landed on a drone ship)`
- Convert Landing Outputs into Training Labels
- [Notebook link to GitHub]



EDA with Data Visualization

- Categorical Plot: FlightNumber vs. PayloadMass
  - How continuous launch attempts and Payload would affect the launch outcome.
- Categorical Plots: FlightNumber vs LaunchSite, Launch Site vs Payload Mass, FlightNumber vs Orbit, Payload vs. Orbit
  - Visualize the relationship between the two variables
- Bar Chart: sucess rate of each Orbit
  - Understand the relationship between success rate of each orbit type
- Line Plot: Year vs Success Rate
  - Get the average launch success trend.
- Notebook Link to GitHub

EDA with SQL

- Unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass.
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Notebook Link to GitHub

Build an Interactive Map with Folium

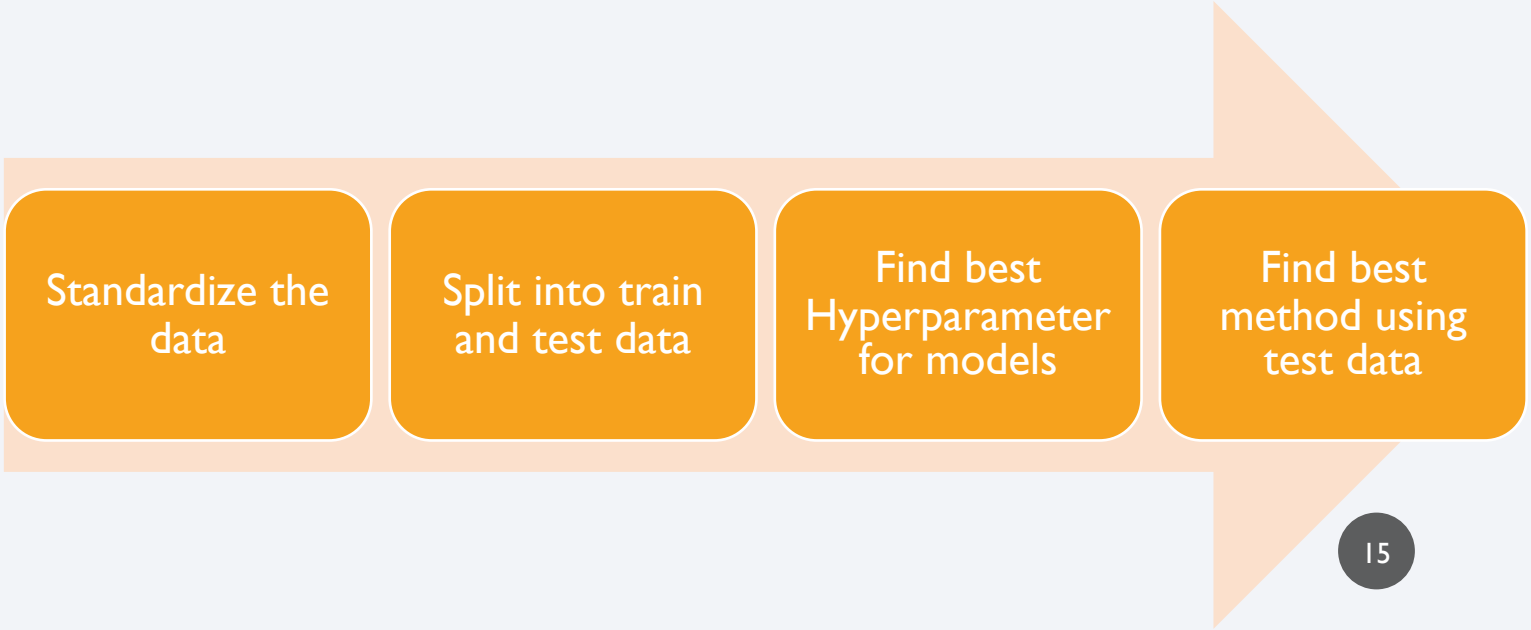
- Circle and Marker at NASA Johnson Space Center's coordinate with a popup label showing its name
  - Add a highlighted circle area with a text label on a specific coordinate.
- Circle and Marker for each launch site
  - Highlight and Locate each launch site on the map, find if they are close to the coast and/or above the equator
- Markers for launch outcomes
  - Mark the success/failed launches for each site on the map
- Line between a launch site to a selected coastline point, railway, highway, or city
  - Calculate the distances between a launch site to its proximities
- Notebook Link to GitHub

Build a Dashboard with Plotly Dash

- Launch Site Drop-down Input Component
  - Select one specific site, check success count and success rate.
- Pie chart visualizing launch success counts.
  - Understand success rate by specific launch site or all
- Range Slider to Select Payload
  - Find if variable payload is correlated to mission outcome
- Scatter plot: Payload vs Launch outcome by Launch Site
  - Visually observe how payload may be correlated with mission outcomes for selected site(s).
- Notebook Link to GitHub

Predictive Analysis (Classification)

- We want to build a classification model that would help us predict if the launch will be successful or not
- We make sure that the features have the same scale, then we split our data into train and test set. We then train the model using cross validation and trying different combination of hyperparameters. The best model and hyperparameter combination will be found when we find the highest accuracy.
- Notebook Link to GitHub

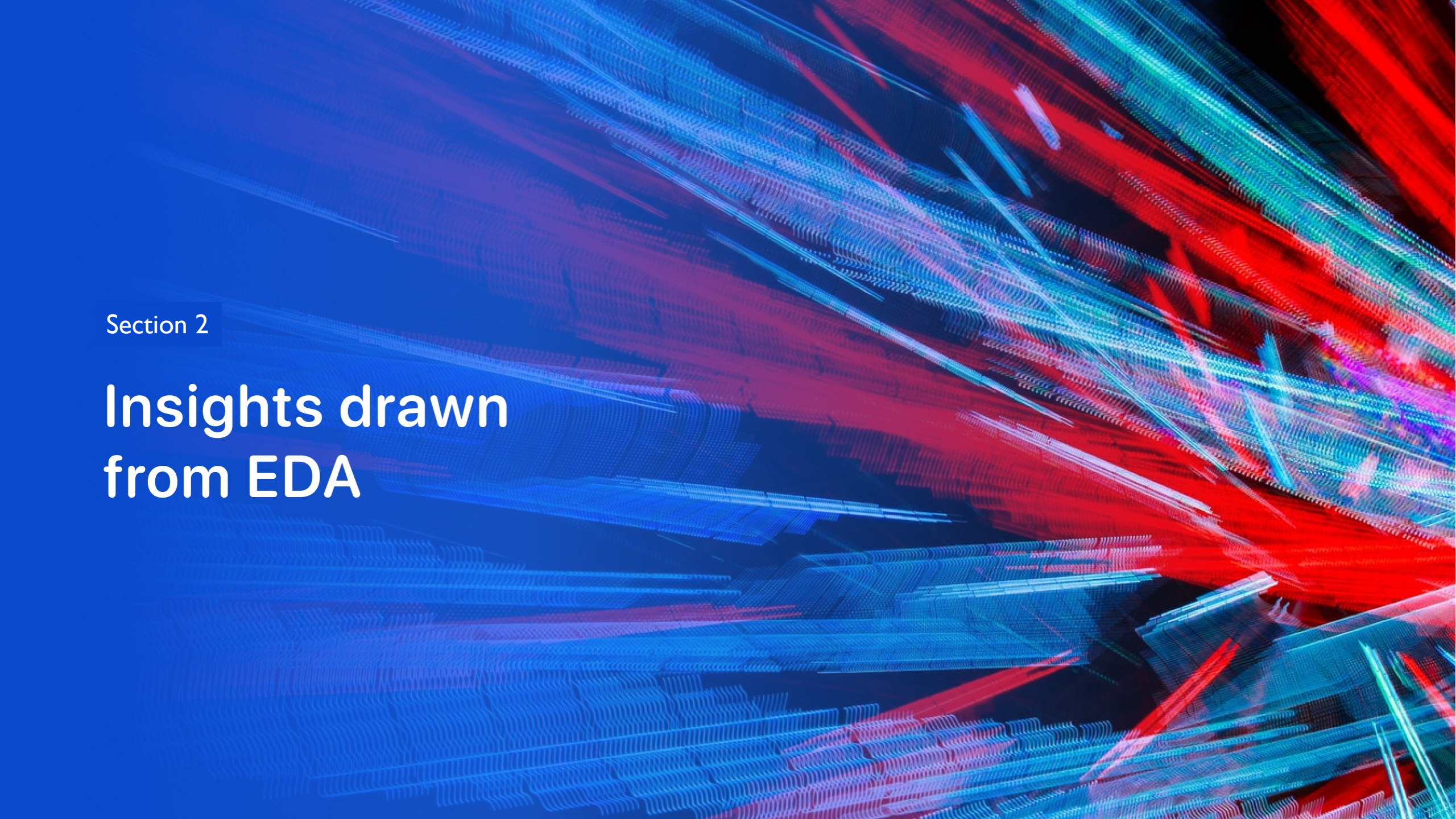


Results

- The more flight numbers, the more successes we have
- SSO, HEO, GEO and ES-L1 orbits have success rate of 100%.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- The success rate since 2013 kept increasing till 2020
- Total payload carried by boosters from NASA is 45596 and the Average payload mass carried by booster version F9 v1.1 is 2928
- All sites are above the equator line and very close to the coast and they are close to railways and highways and coastlines.

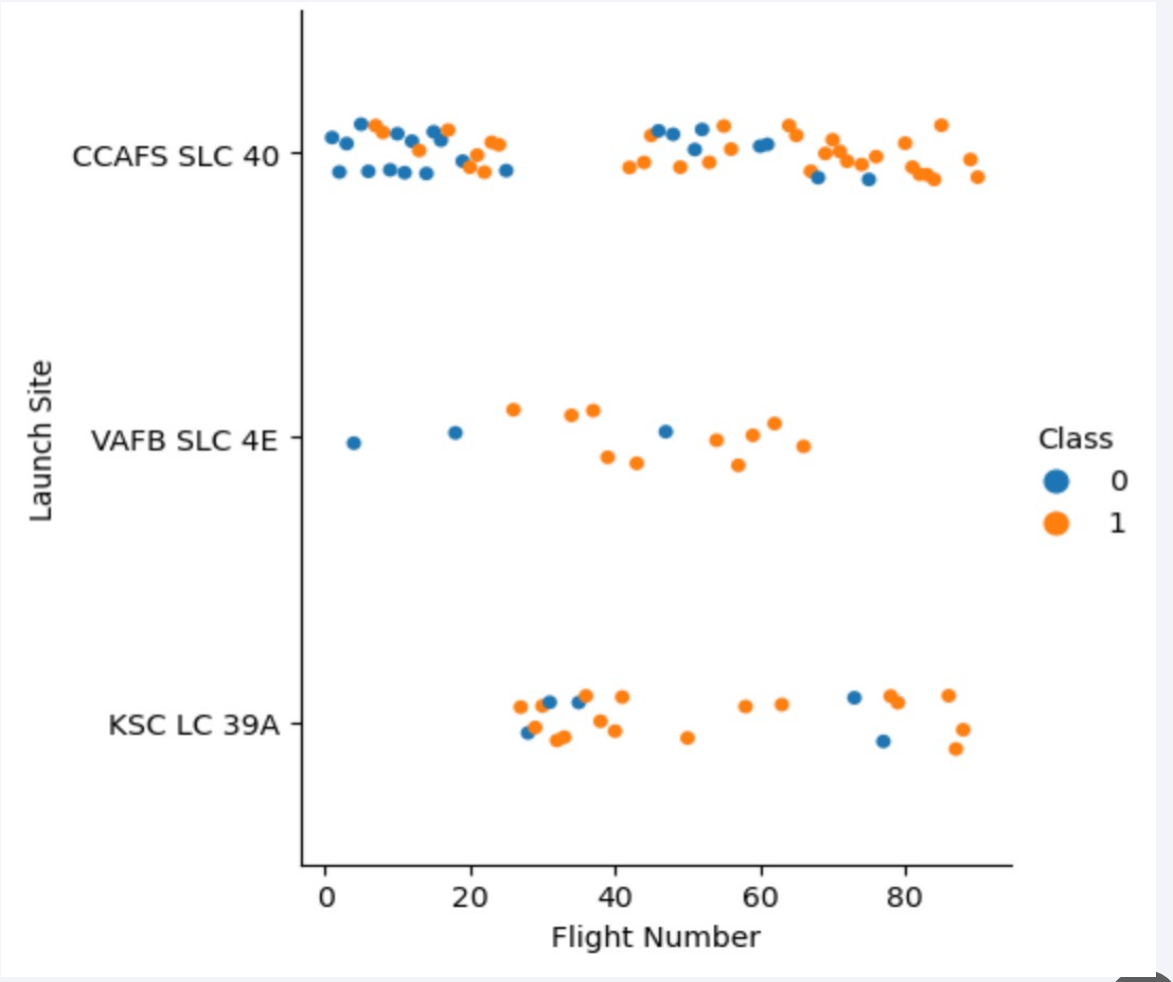
Results

- We have four different launch sites. The site that has the largest successful launches is KSC LC-39A
- KSC LC 39A has the highest launch success rate with 76.9 % of launches being successful.
- The payload range that has the highest launch success rate is 2000-4000 and 4000 - 9000 has the lowest payload launch success rate
- The F9 Booster version that has the highest launch success rate is FT
- The KNN model has the highest accuracy with 0.84 for train and 0.94 for test. Only 1 False positive


Section 2
Insights drawn from EDA

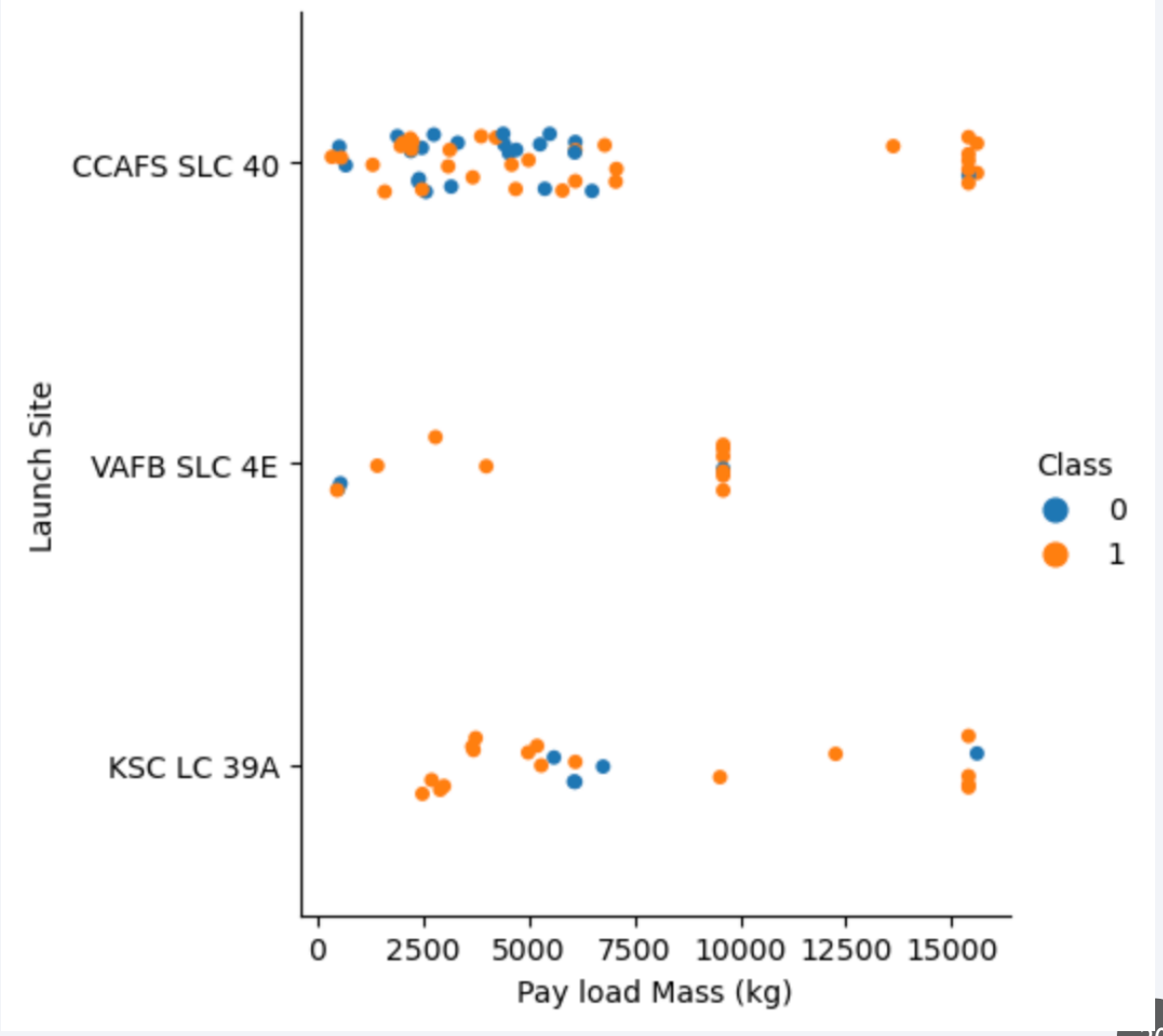
Flight Number vs. Launch Site

- The more flight numbers, the more successes we have
- CCAFS LC-40 has the greatest number of flights but also the greatest number of fails and successes.
- When the number of flights is beyond 77, the success rate is 100% for CCAFS LC-40.
- The 77% rate of the other two outcomes is noticeable in the graph.



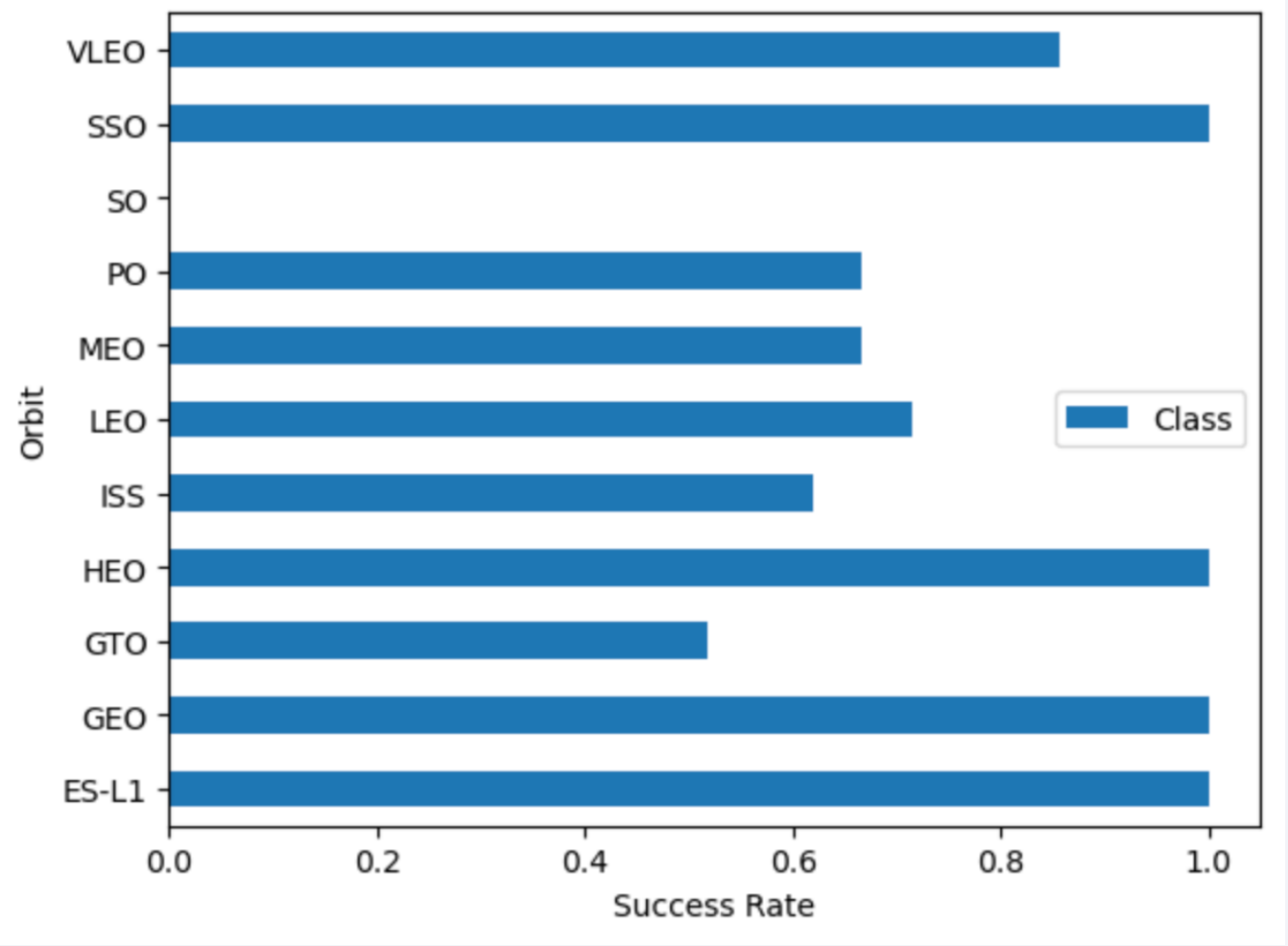
Payload vs. Launch Site

- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



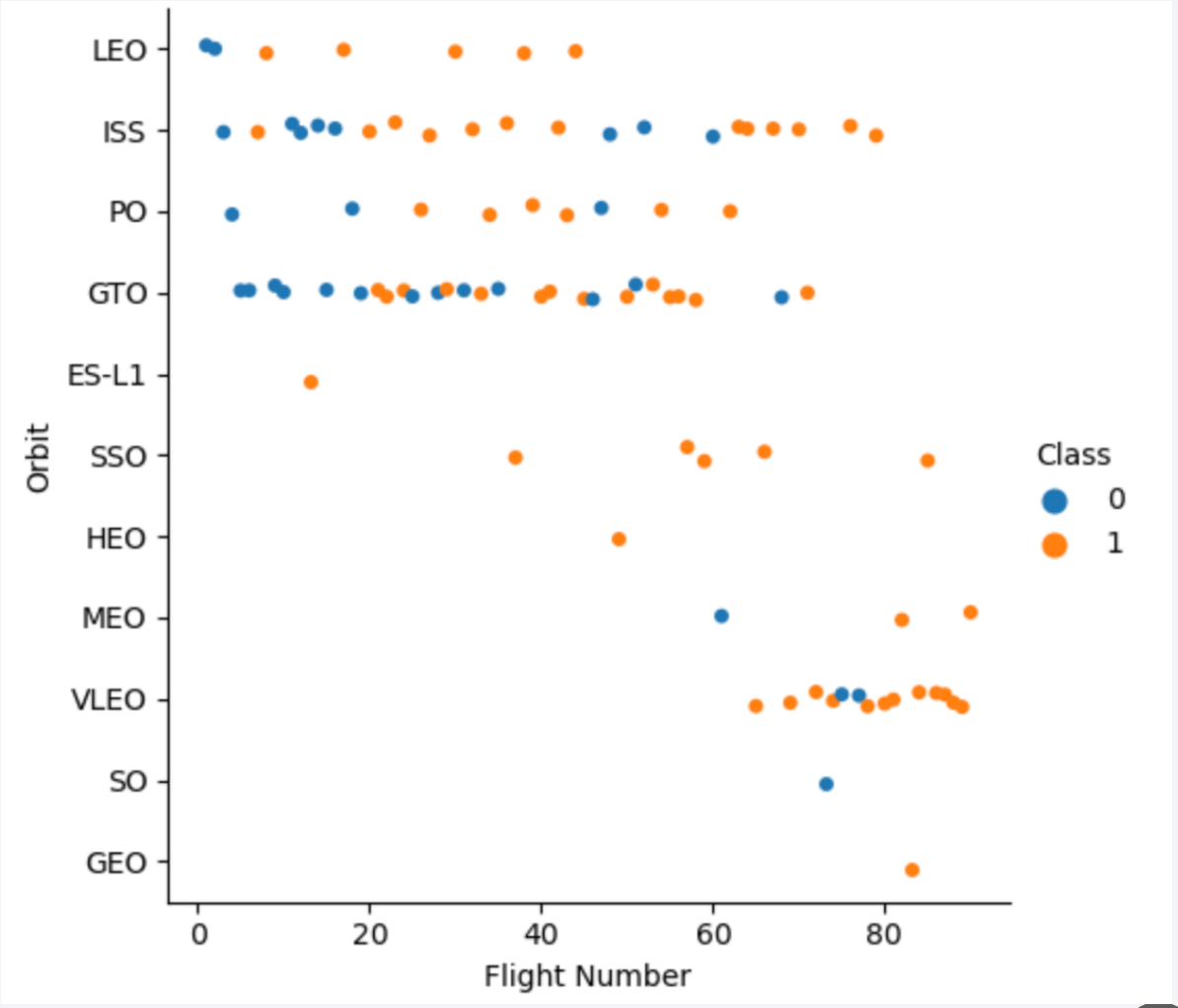
Success Rate vs. Orbit Type

- SSO, HEO, GEO and ES-L1 have success rate of 100%.
- In the other hand, SO has a success rate of 0%.
- VLEO has a success rate greater than 80%



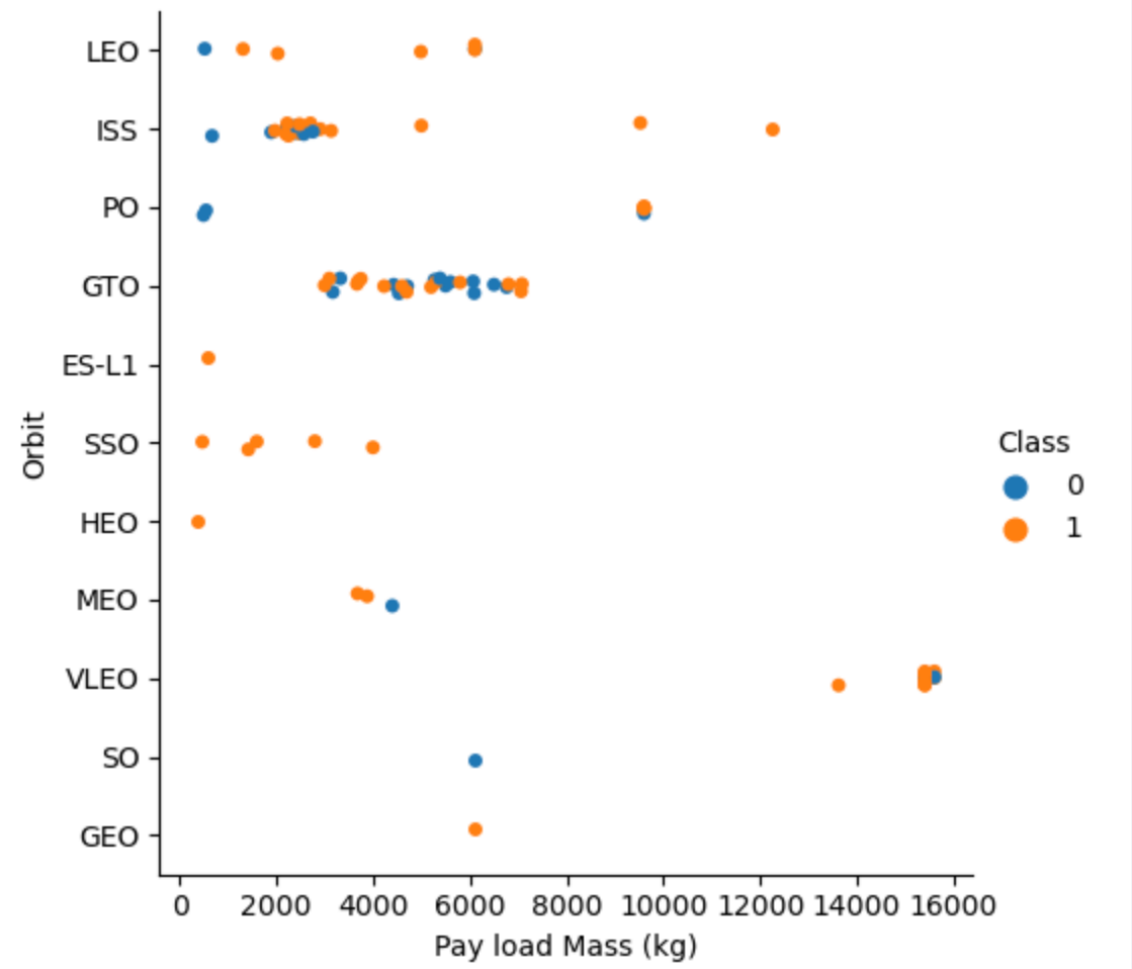
Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights;
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.



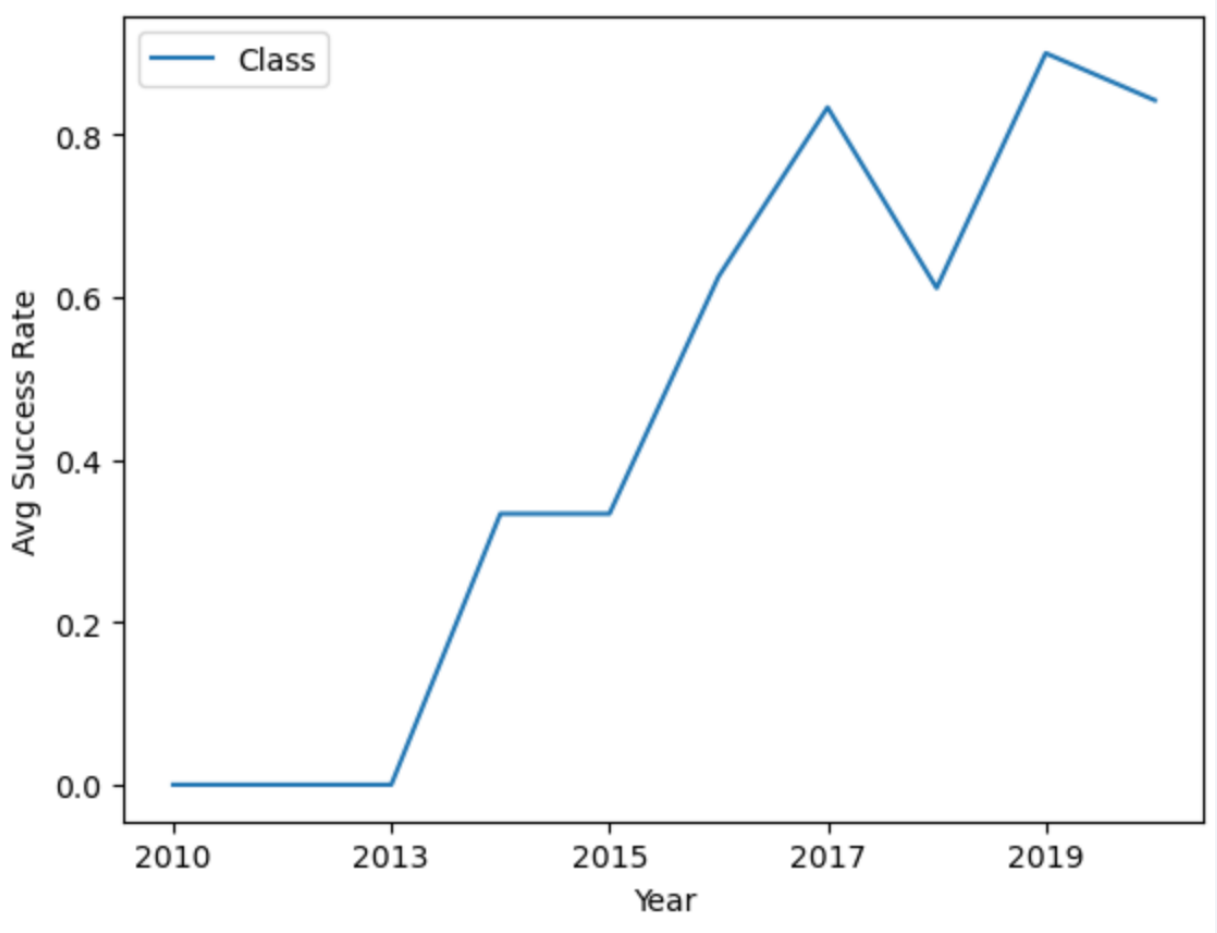
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

- There are four unique Launch Sites names in the space mission:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Records where launch sites begin with the string 'CCA'

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA

customer	total_payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

booster_version	avg_payload_mass
F9 v1.1	2928

First Successful Ground Landing Date

- Date when the first successful landing outcome in ground pad was achieved.

1
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

booster_version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Landing _Outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

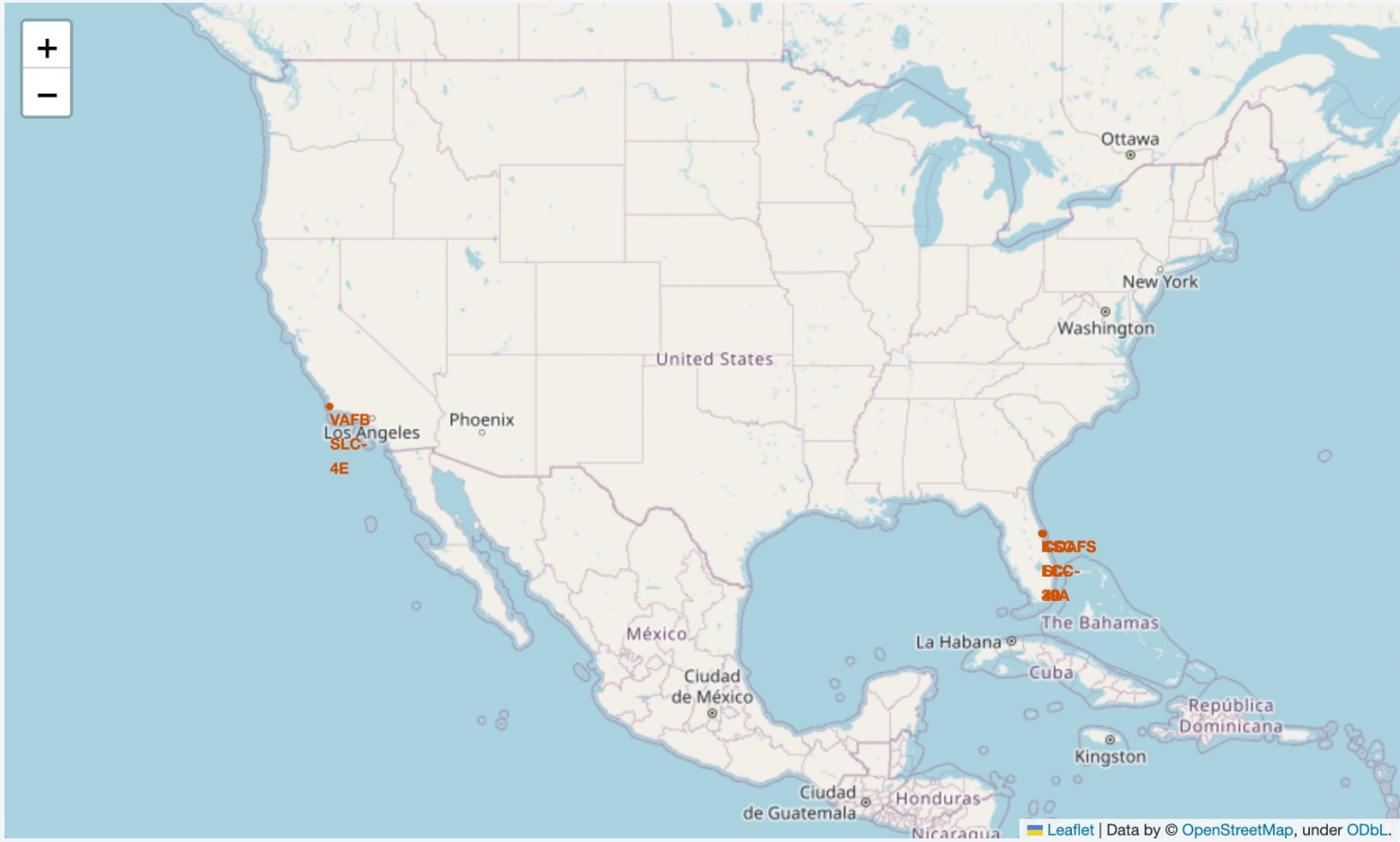
Landing _Outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

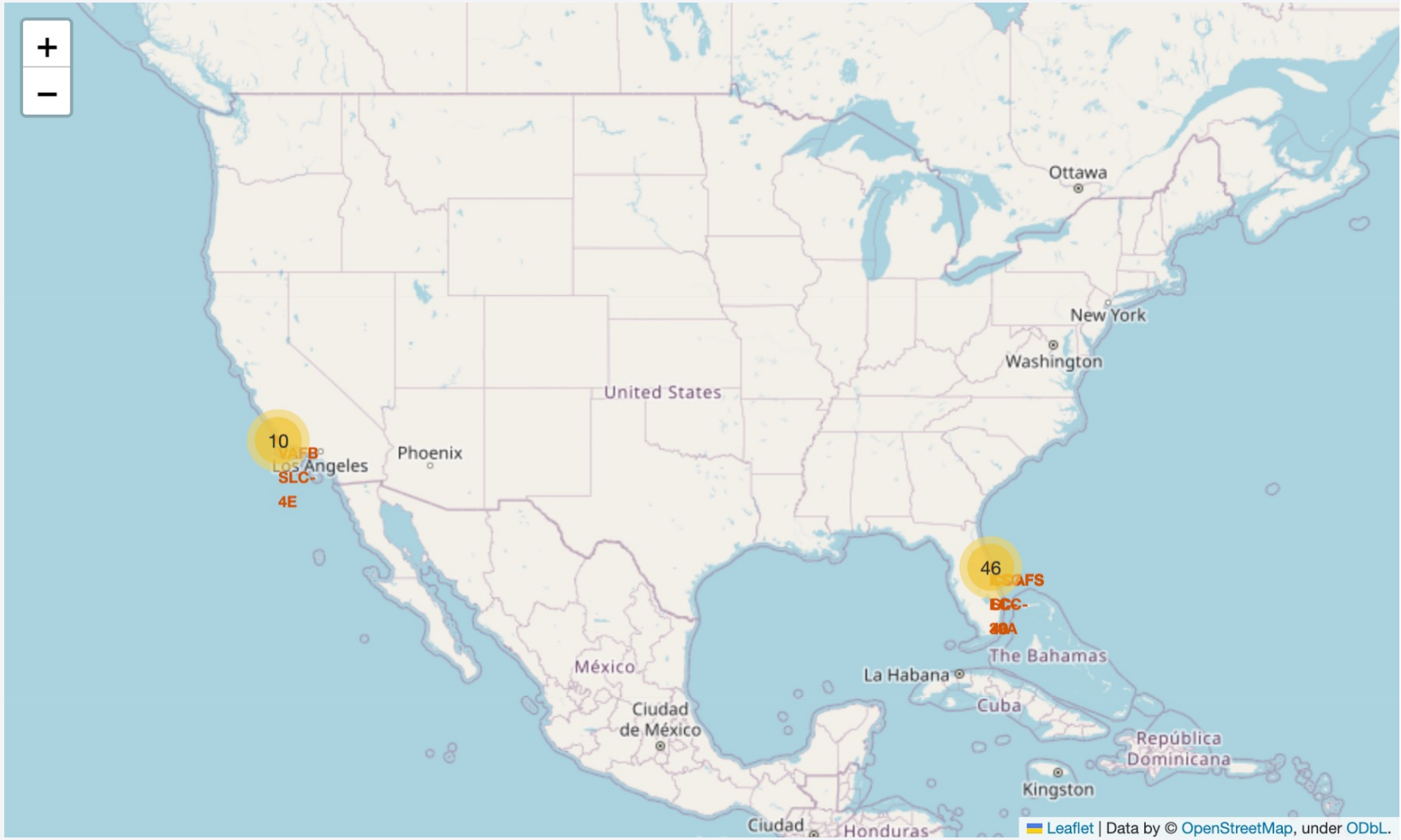
Launch Sites on map

- All sites are above the equator line
- All sites are very close to the coast



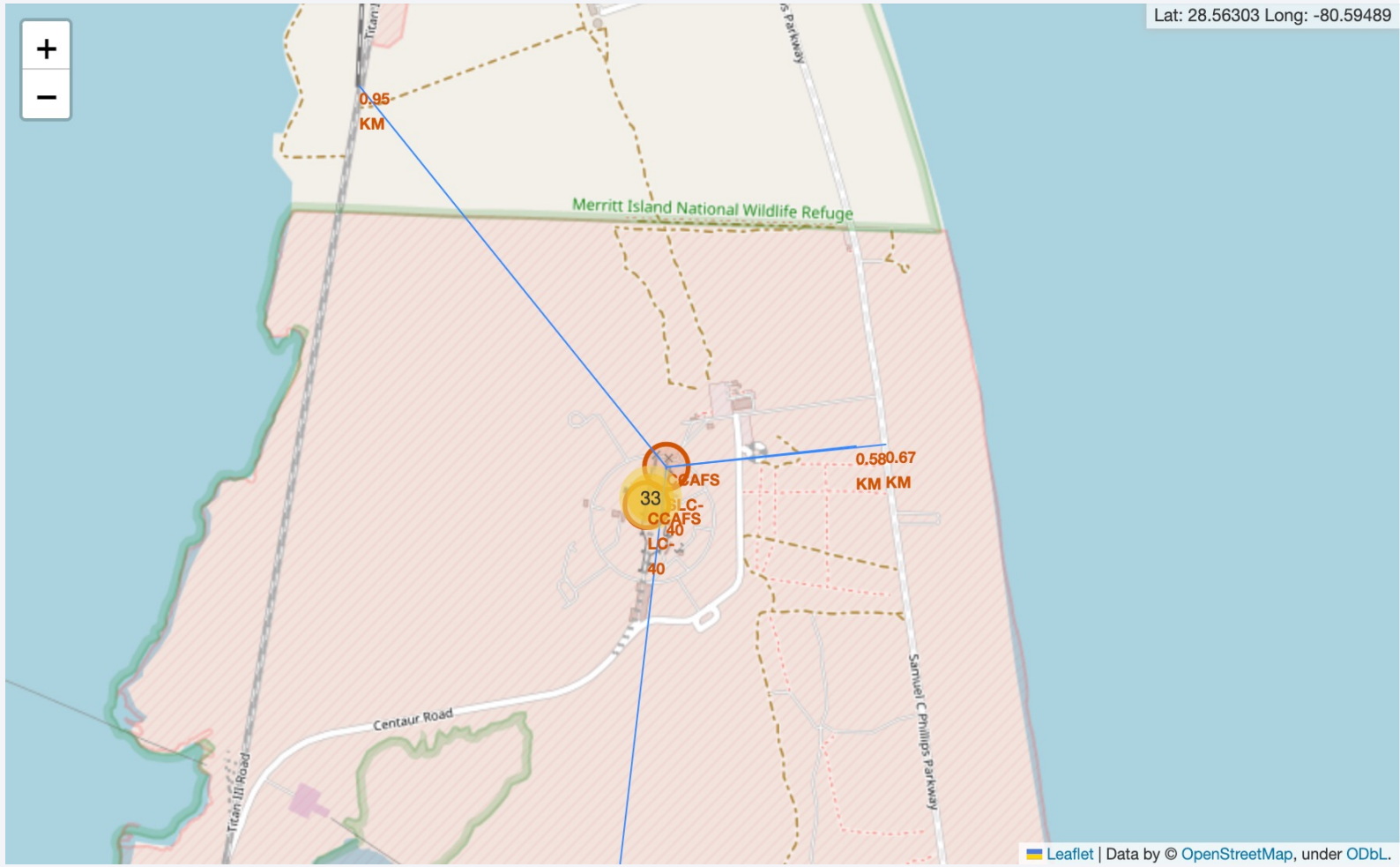
Success/Failed launches on the map

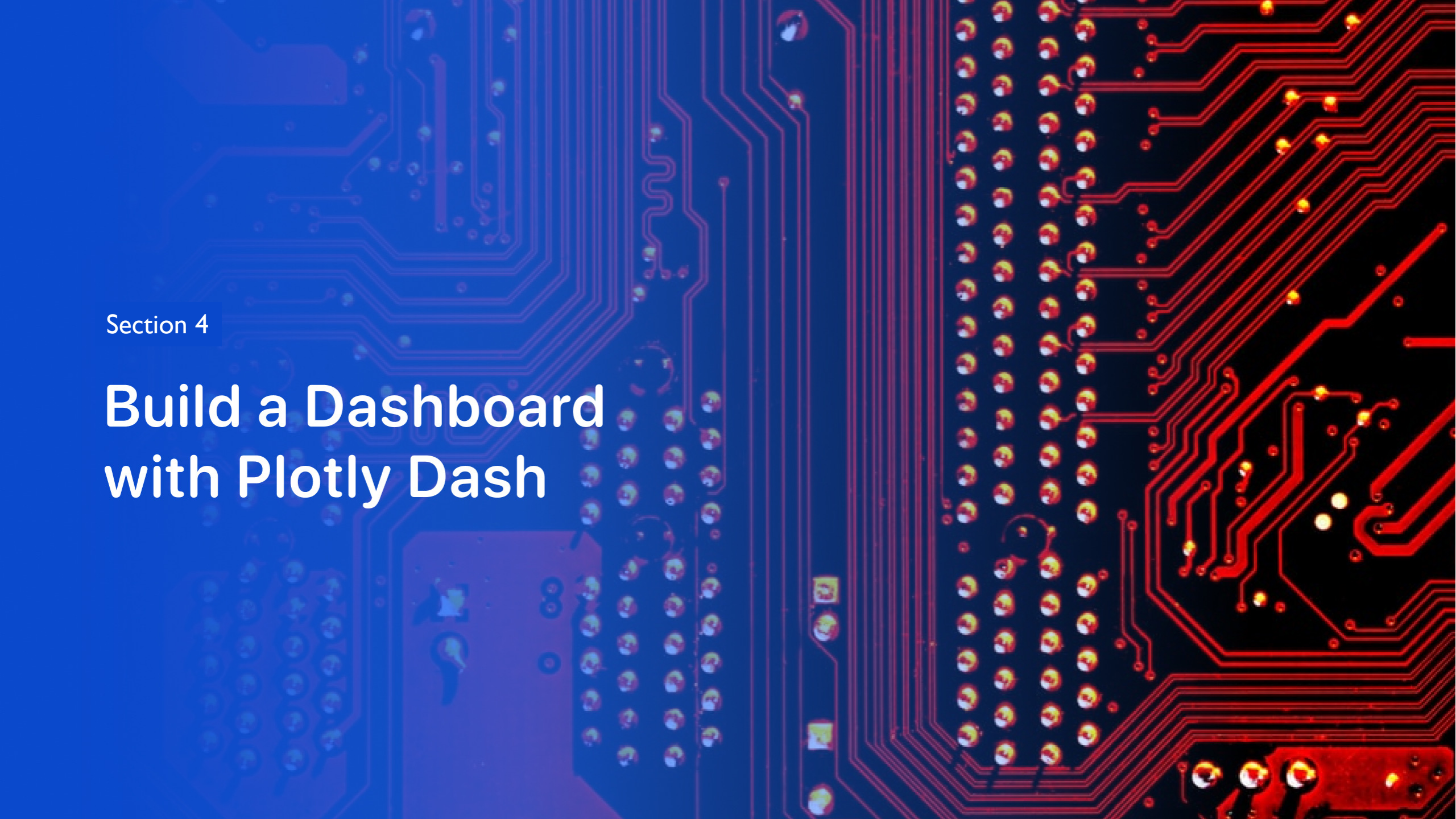
- There are 10 launches on the west coast and 46 on the east coast
- VAFB SLC-4E on the west coast has 4 successful launches out of 10
- KSC LC-39A has 10 successful launches out of 13
- CCAFS LC-40 has 7 successful launches out of 26
- CCAFS SLC-40 has 3 successful launches out of 7



Launch site to its proximities

- Launch sites are close to railways and highways and coastlines.

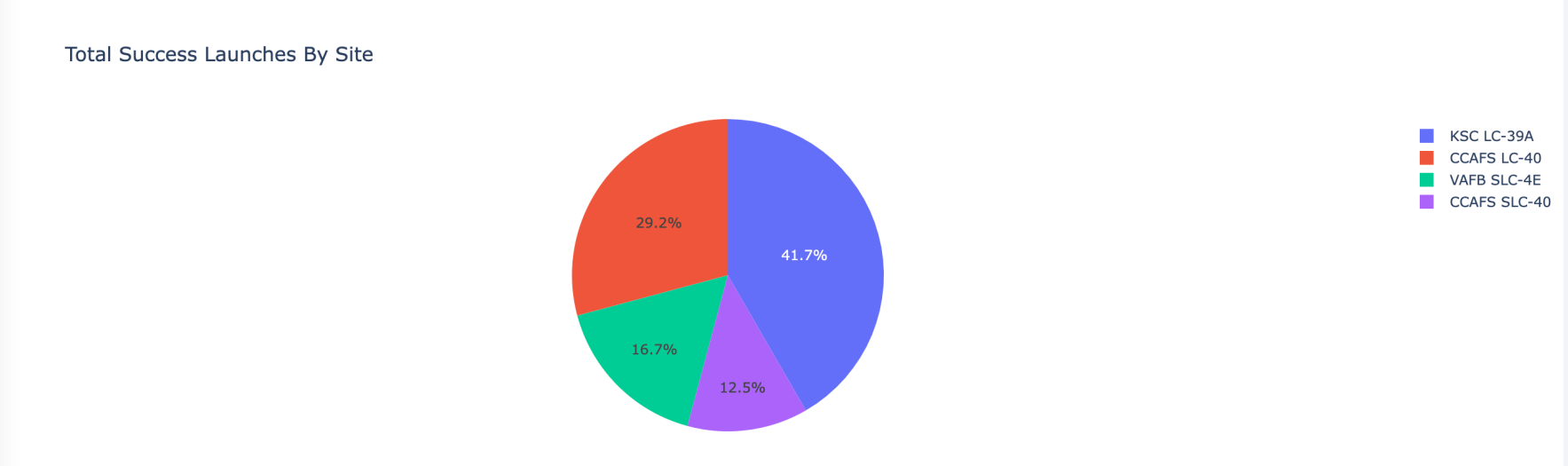




Section 4

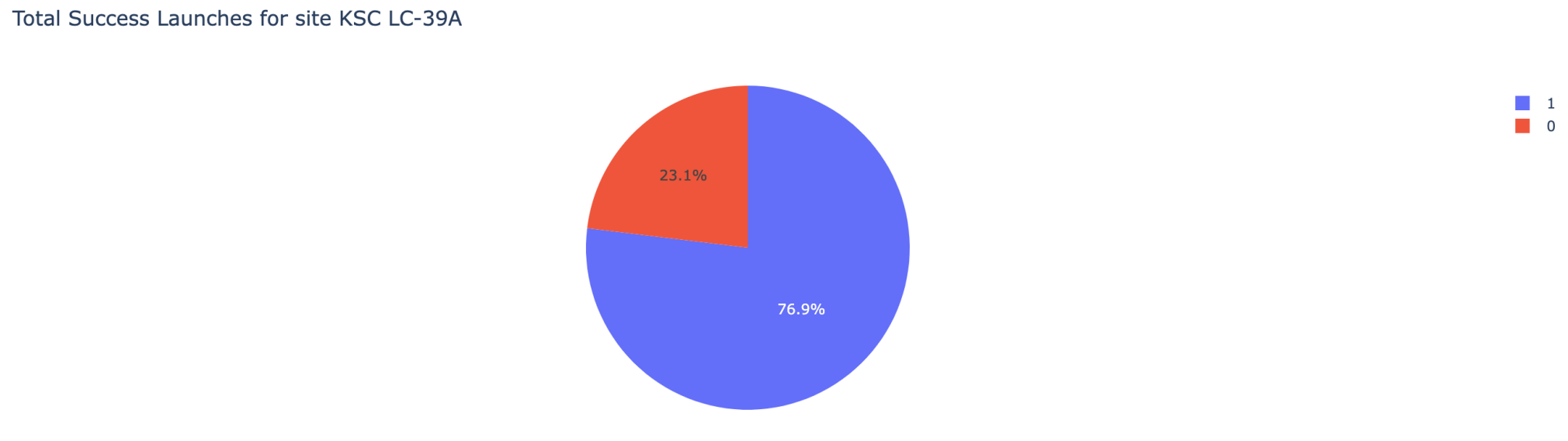
Build a Dashboard with Plotly Dash

Launch Site Drop-down Input Component



- We have four different launch sites. The site that has the largest successful launches is KSC LC-39A

Launch site with highest launch success ratio



- KSC LC 39A has the highest launch success rate with 76.9 % of launches being successful.

Payload range(s) vs Launch Outcome

- The payload range that has the highest launch success rate is 2000-4000
- 4000 - 9000 has the lowest payload launch success rate

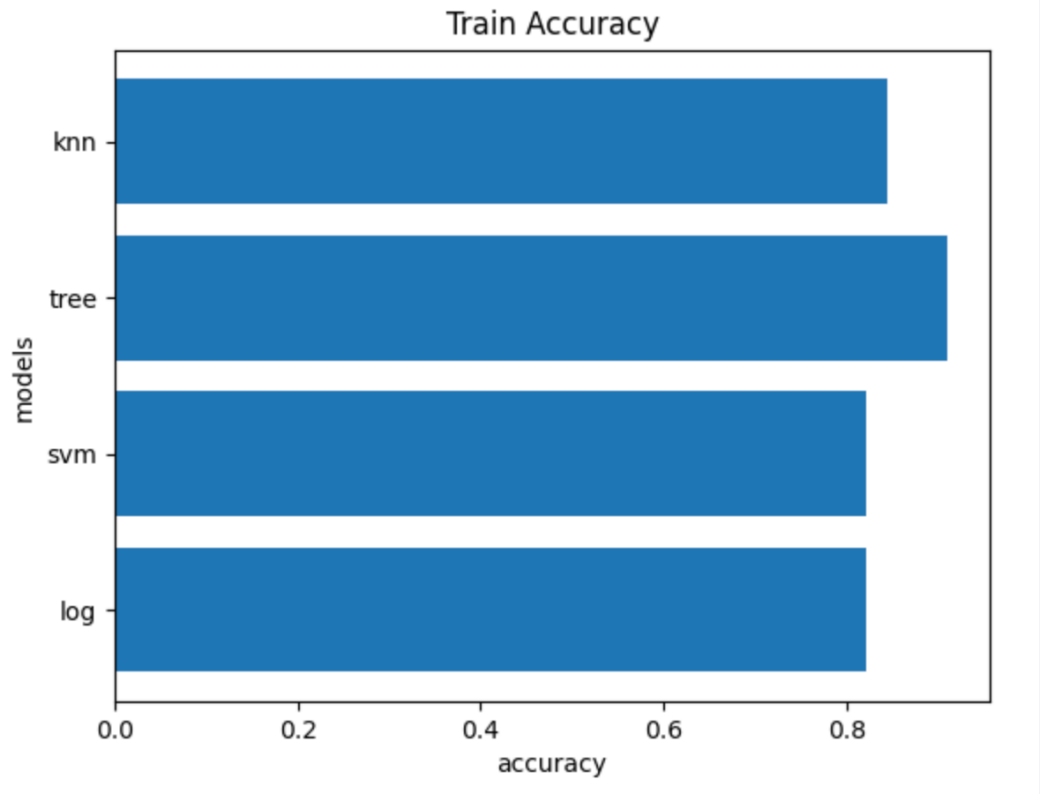
Payload range(s) vs Launch Outcome

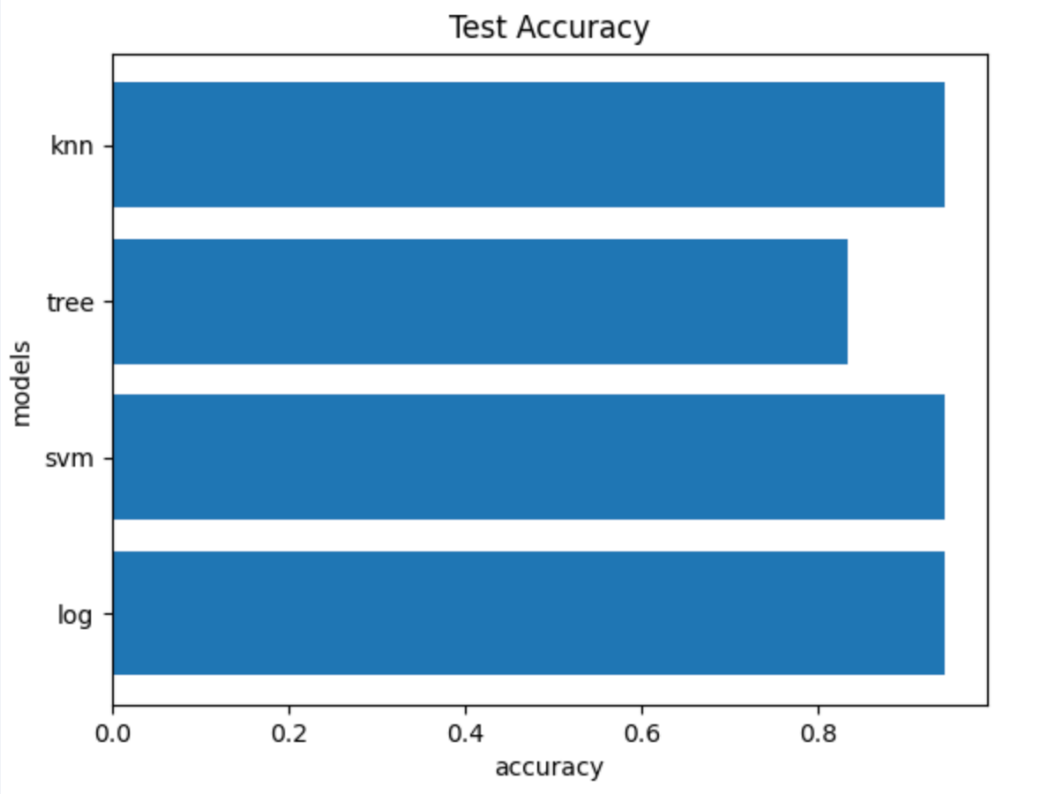
- The F9 Booster version that has the highest launch success rate is FT

Section 5

Predictive Analysis (Classification)

Classification Accuracy

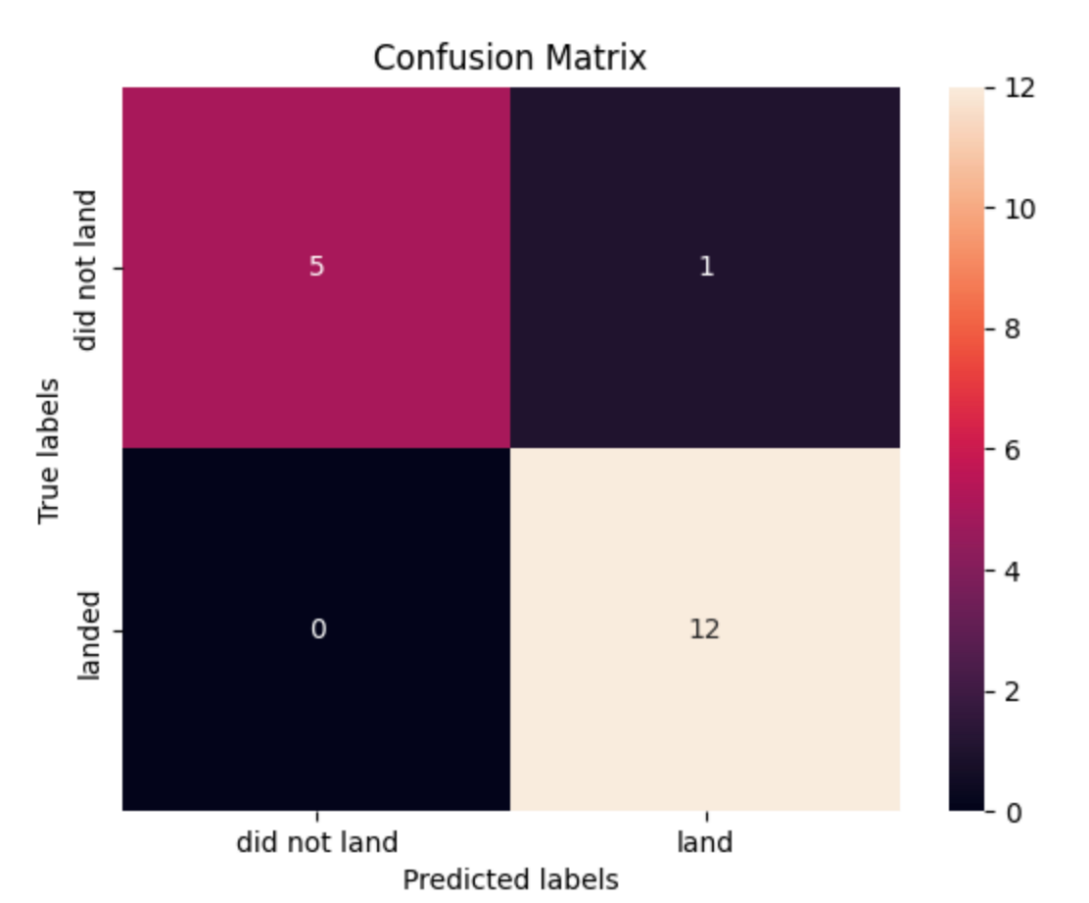




- When using the tree model, we obtained the highest training accuracy. Also, the highest testing accuracy by a small amount is the knn model. While it is true that knn does not have the highest train accuracy, it is actually generalizing better, so that is our best model.

Confusion Matrix

- Examining the confusion matrix, we see that the knn model can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- Launch sites are close to the coast and railways but far away from cities
- The highest launch success rate is 76.9 %
- The payload range that has the highest launch success rate is 2000-4000
- With our model, we can predict if the next landing outcome will be successful or not
- The KNN model has the highest accuracy with 0.84 for train and 0.94 for test. Only 1 False positive
- We can still improve our model with more data, maybe more features, more up-to-date data
- We can try neural networks next time or other combination of hyperparameters with the same models

Appendix

- Web Scraping dataset

```
df = pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

```
df.head()
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0003.1	Failure	4 June 2010	18:45
2	2	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0004.1	No attempt\n	8 December 2010	15:43
3	3	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0005.1	No attempt	22 May 2012	07:44
4	4	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt\n	8 October 2012	00:35

Appendix

- Landing Outcomes

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

Appendix

- SQL queries

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

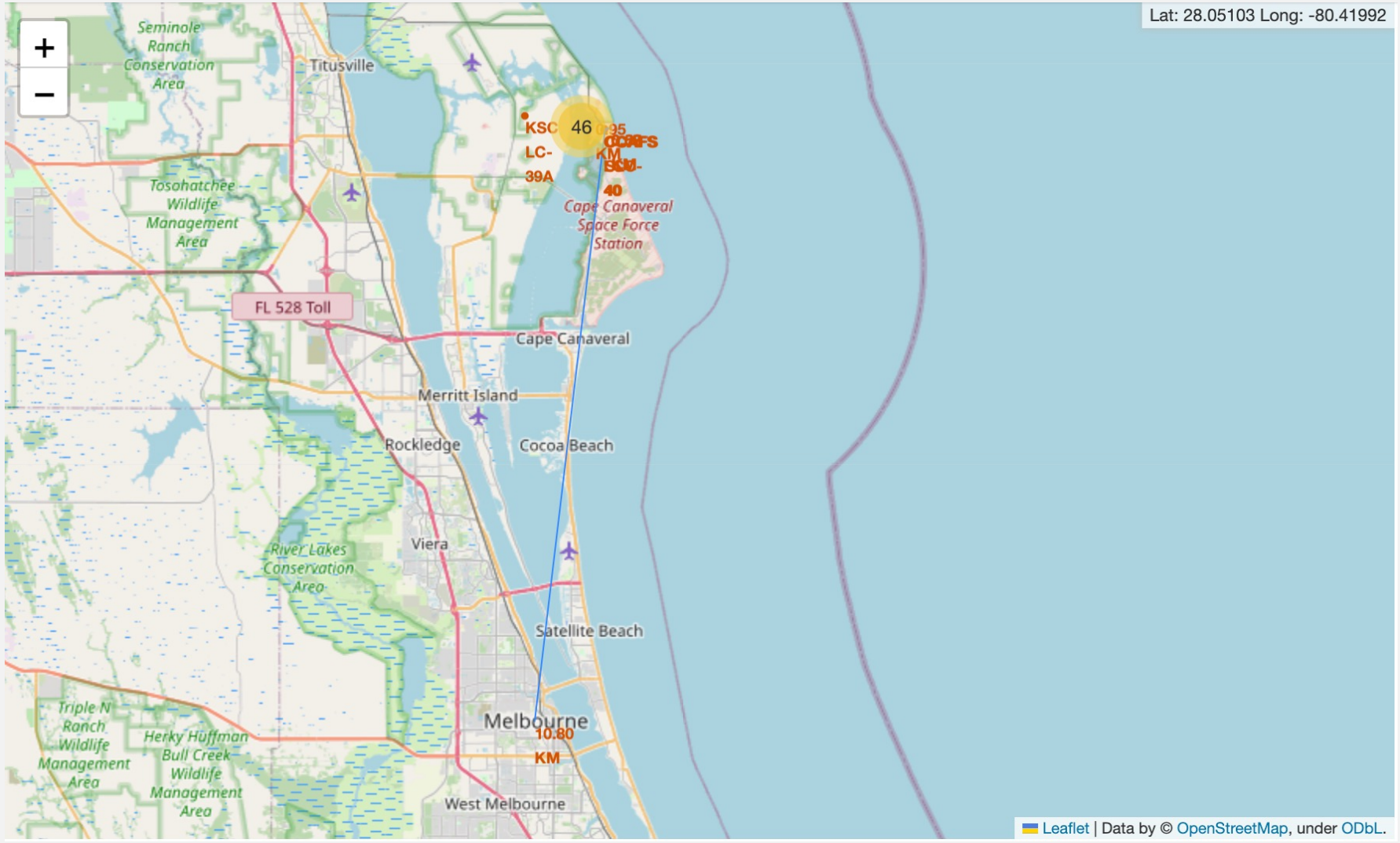
```
%%sql
SELECT "Landing _Outcome", COUNT(*) FROM SPACEX
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing _Outcome"
ORDER BY COUNT(*) DESC
```

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT "Landing _Outcome", booster_version, launch_site, DATE FROM SPACEX
WHERE YEAR(DATE)=2015 AND "Landing _Outcome" = 'Failure (drone ship)'
```

Appendix

- Distance From Launch Site to the closest city: Melbourne



Appendix

- KNN training code

Create a k nearest neighbors object then create a `GridSearchCV` object `knn_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters`.

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
```

```
knn_cv = GridSearchCV(KNN, parameters, cv=10)
knn_cv.fit(X, Y)
```

```
GridSearchCV(cv=10, estimator=KNeighborsClassifier(),
             param_grid={'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
                         'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
                         'p': [1, 2]})
```

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 5, 'p': 1}
accuracy : 0.8444444444444444
```

Thank you!